Обучение генеративных нейросетевых моделей с использованием технологий распределенных вычислений

2019, Май, МИЭМ

Андрей Устюжанин

НИУ ВШЭ

ICL

Генеративные модели

Цели

- воспроизведение известных условных распределений

Специфика обучения

- сложная функция
- большие наборы данных,
- длительное время обучения

Generative models [edit]

Types of generative models are:

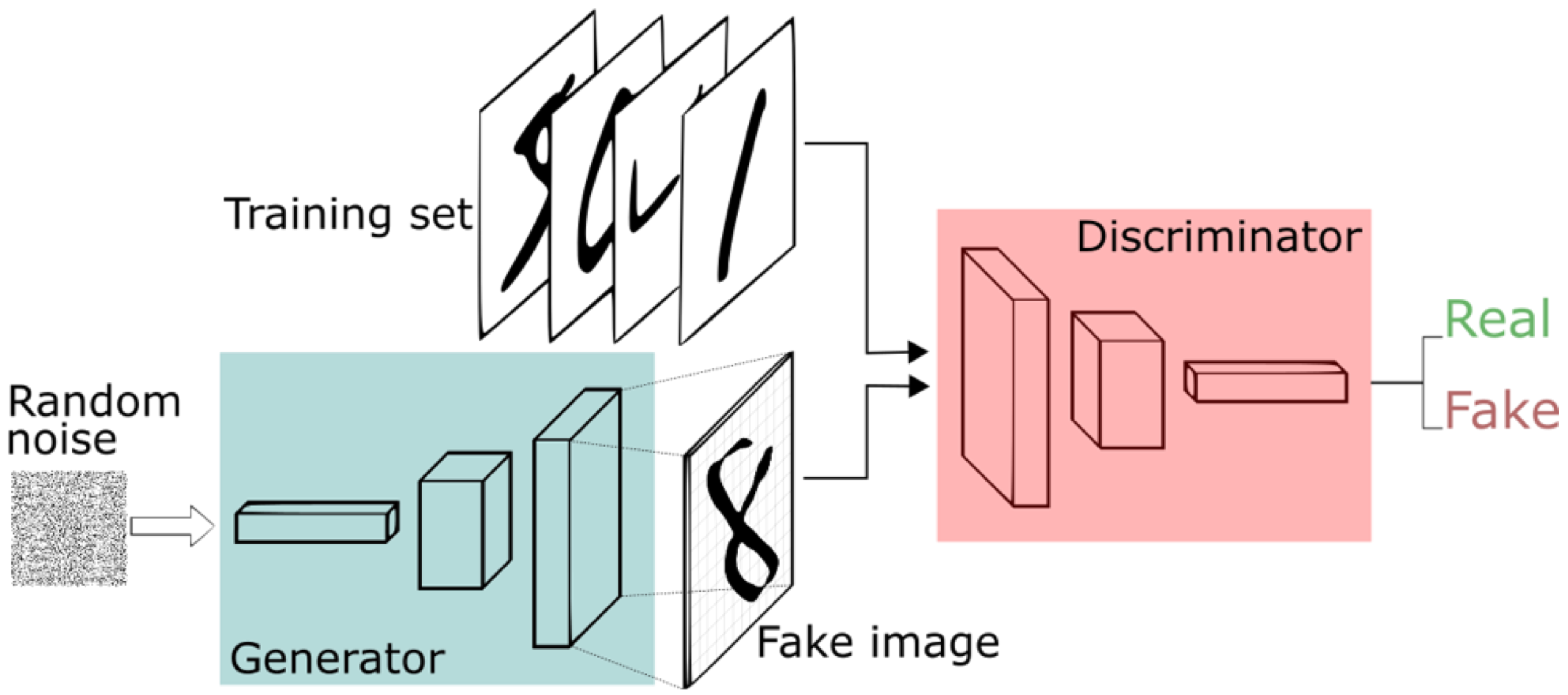
- Gaussian mixture model (and other types of mixture model)
- Hidden Markov model
- Probabilistic context-free grammar
- Bayesian network (e.g. Naive bayes, Autoregressive model)
- Averaged one-dependence estimators
- Latent Dirichlet allocation
- Boltzmann machine (e.g. Restricted Boltzmann machine, Deep belief network)
- Variational autoencoder
- Generative adversarial network
- Flow-based generative model

WIKIPEDIA
The Free Encyclopedia

Генеративные модели

Генеративные состязательные сети

- Обучение 2 сетей одновременно



https://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf

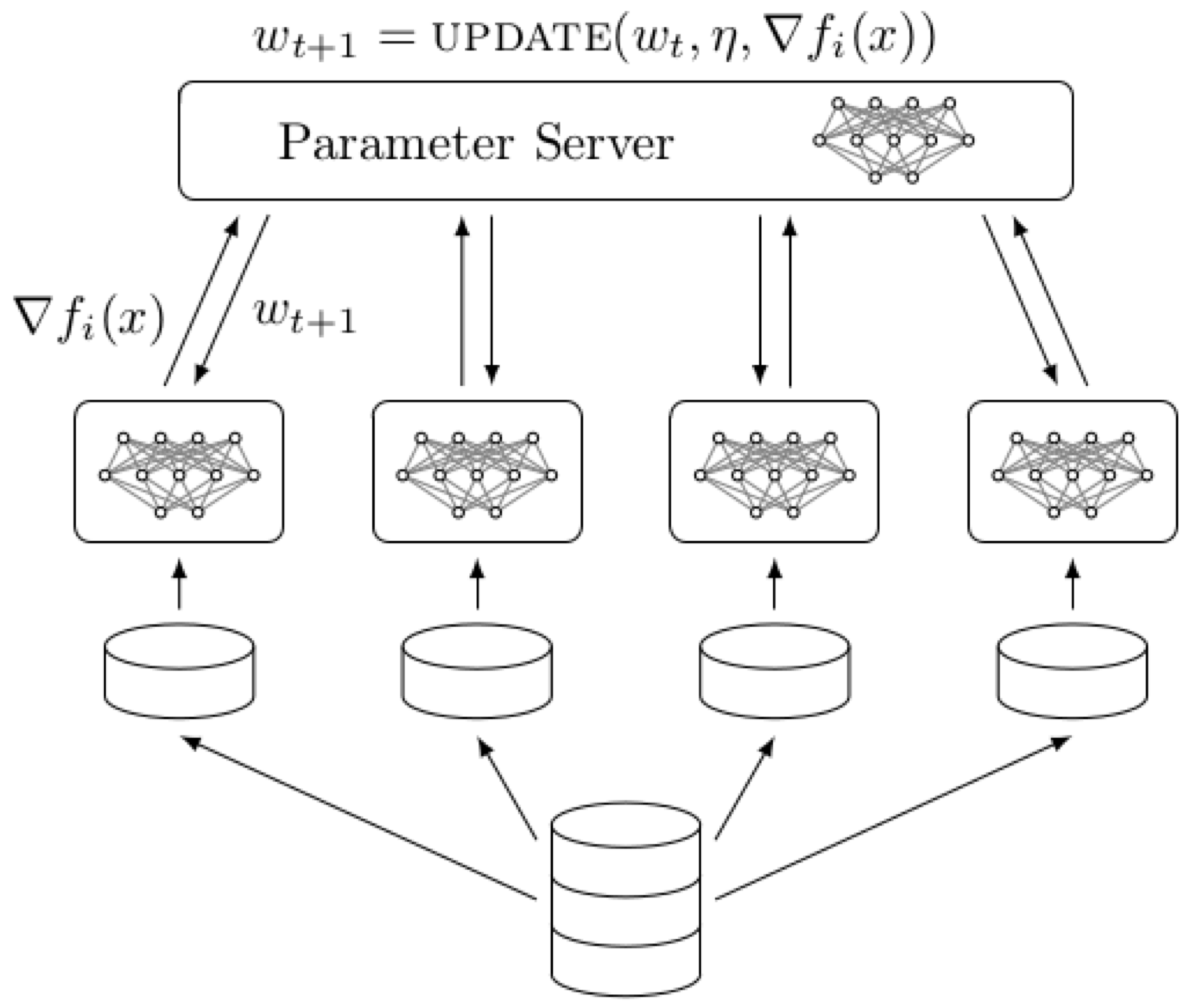
Пример

Механизмы обучения сложных моделей

Нативные возможности библиотек

tensorflow.distributed, torch.distributed
https://github.com/tmulc18/Distributed-TensorFlow-Guide, https://pytorch.org/docs/stable/distributed.html

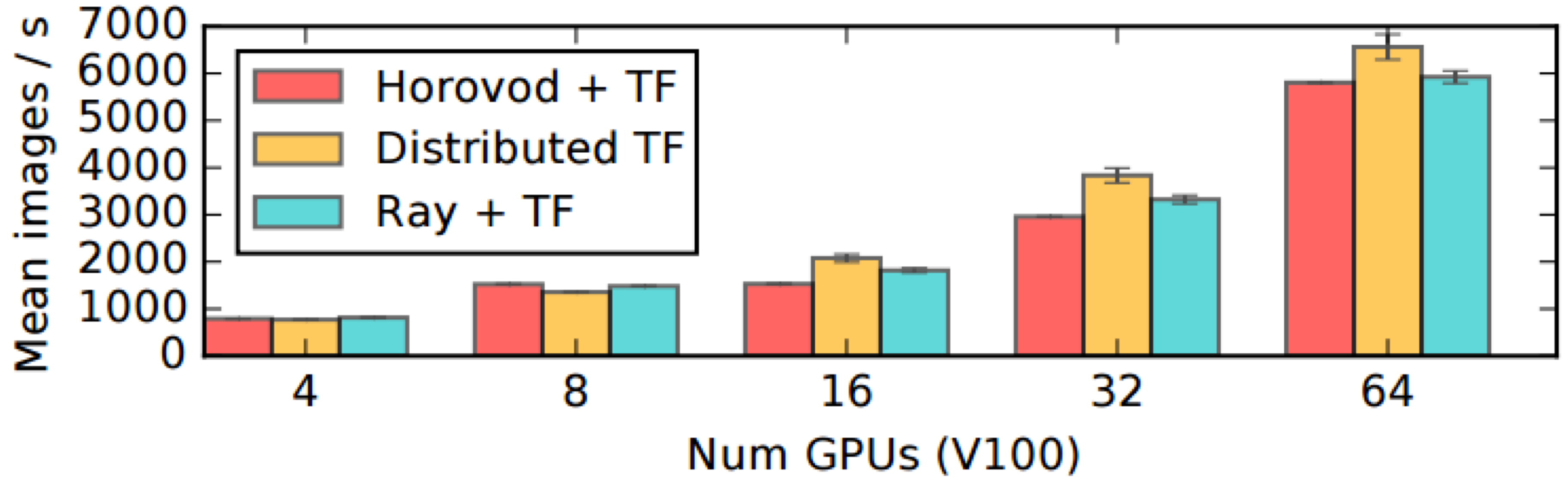
› Сложно готовить, позволяют больше гибкости



Высокоуровневые обертки

Horovod, https://github.com/horovod/horovod
Ray, https://github.com/ray-project/ray

› Удобны в использовании

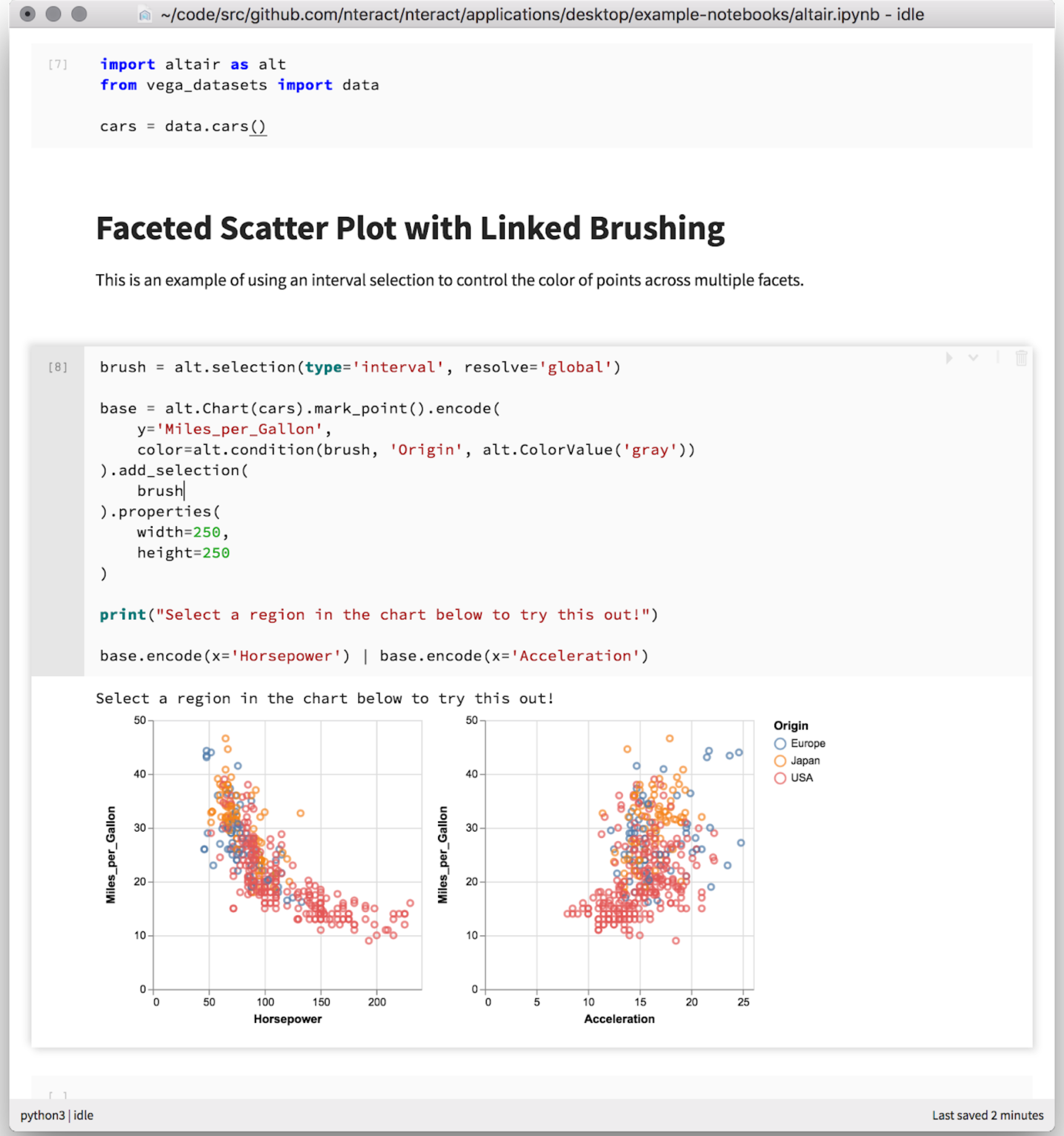


Исследовательский процесс

1. Формулировка задачи
2. Повторить 3-5 до получения результата:
3. Выбор модели
4. Анализ результатов, сравнение с аналогами
5. Уточнение постановки задачи
6. Публикация статьи (репозиторий с исходным кодом)

Инструменты для разработки

- Быстрые эксперименты, Jupyter, Jupyterlab
  - Nbviewer, reviewnb (https://www.reviewnb.com/) и google colab
  - https://www.kth.se/blogs/pdc/2019/01/using-jupyter-notebooks-to-manage-slurm-jobs/
- Версионирование
  - кода(gitlab, github)
  - окружений (pip, conda, docker)
  - наборы данных (?)
- Подбор гиперпараметров
  - Ray (tuneML), scikit-optimize



Контроль разработки

Мониторинг качества моделей

- comet.ml, https://github.com/Atmyre/LHCb_PID_Compression/blob/master/exp_two-head-gan/gan_torch-5dim-cometml-delta1.ipynb
- https://www.comet.ml/atmyre/gan-torch-5dim-cometml-delta1-4/c156ce27e5d24b84956ceba90241f8ad/images

Автоматическая проверка (Continuous Integration)

- https://medium.com/netflix-techblog/scheduling-notebooks-348e6c14cfd6

Вычислительные ресурсы лаборатории

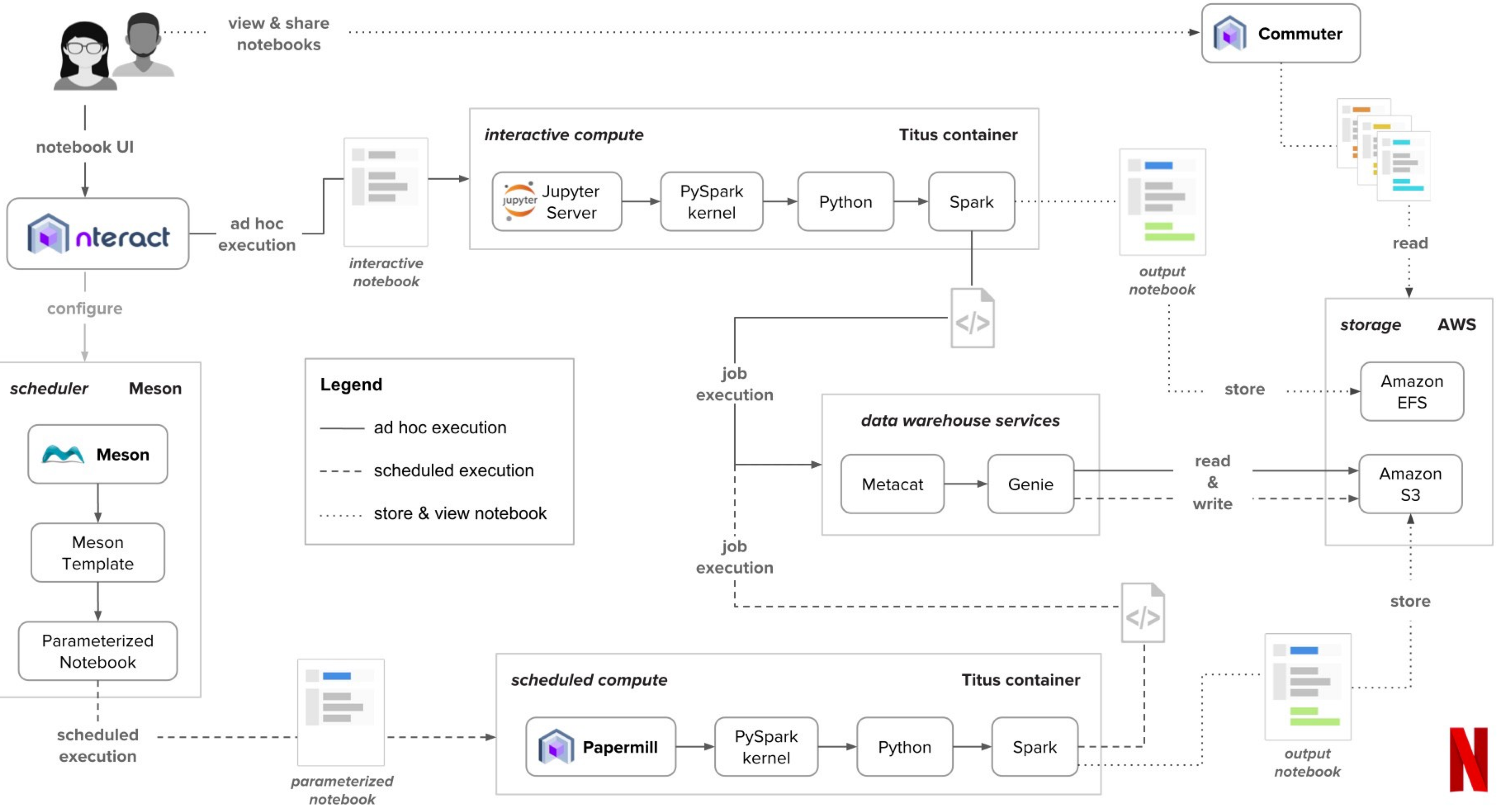
Azure

- Pre-built virtual machines
- Azure Container Service
- Azure Kubernetes Service

Гибкость настройки окружений

Интеграция с сервисами

Netflix example



https://medium.com/netflix-techblog/notebook-innovation-591ee3221233

Пожелания по работе с кластером НИУ ВШЭ

Узлы кластера

- Управление окружением Singularity, conda (+)
- Снять ограничение на исходящий трафик с узлов в Internet

Задачи

- Могут выполняться до 10 дней
- Могут занимать от 0.25 GPU до 16 GPU
- Обучение нескольких моделей параллельно

Дополнительное ПО / Сервисы

- интеграция slurm с gitlab (https://gitlab.com/gitlab-org/gitlab-runner/merge_requests/667)
- https://github.com/nteract/commuter

Заключение

1. Решение фундаментальных и прикладных задач машинного интеллекта требует не только адекватных вычислительных мощностей, но и адекватных механизмов для

> **эффективного распределения задач по ресурсам,**
> **работы с окружениями,**
> **контроля версий,**
> **автоматизации,**
> **мониторинга**

2. Сделать нормы (policy) более открытыми, а балансировку обеспечит биллинг

Ссылки

- Jupyter Notebooks managed by slurm https://cloud.google.com/community/tutorials/using-slurm-to-host-jupyter-notebooks-review
- https://github.com/SiLiKhon/RichGANatACAT2019
- https://medium.com/netflix-techblog/notebook-innovation-591ee3221233
- Scheduling Notebooks at Netflix https://medium.com/netflix-techblog/scheduling-notebooks-348e6c14cfd6
- FAIR data principles https://www.force11.org/group/fairgroup/fairprinciples